Multiple Regression Model

INSR 260, Spring 2009
Bob Stine

Overview

- Multiple Regression Model (MRM)
- Estimators, terminology   similar to SRM
- Assumptions   new plots
- Inference   new test
- Prediction   similar to SRM
- Examples   (from Bowerman, Ch 4)
  - Fuel consumption
  - Sales management

Multiple Regression Model

- Equation has **k** explanatory variables
  Mean $E\ Y|X = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k = \mu_{y|x}$
  Observations $y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i$
- Assumptions (as in SRM)
  - Independent observations
  - Equal variance σ^2
  - Normal distribution around "line"
    $y_i \sim N(\mu_{y|x}, \sigma^2)$ $\varepsilon_i \sim N(0, \sigma^2)$
- k+2 parameters identify model
  $\beta_0, \beta_1, \ldots, \beta_k, \sigma^2$

Least Squares

- Criterion
  - Find estimates that minimize sum of squared deviations

    $\min_a \Sigma(y_i - a_0 - a_1 x_{i1} - \ldots - a_k x_{ik})^2$

- Fitted values, residuals
  - Fitted values (on the line) $\hat{y} = b_0 + b_1 x_{i1} + \ldots + b_k x_{ik}$
  - Residual deviations $e = y - \hat{y}$

- Standard error of regression (estimate of σ^2)
  - $s^2 = \Sigma e_i^2/(n-k-1)$
  - degrees of freedom
  - RMSE = square root of s^2

Goodness of Fit

- R-squared statistic
  - Square of correlation between Y and $\hat{Y}$
  - Percentage of "explained" variation
  - Always increases as variables are added to equation

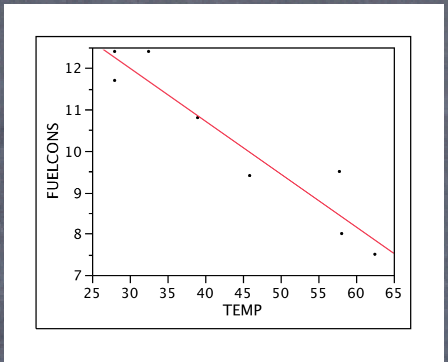
$$R^2 = \frac{\text{Explained SS}}{\text{Total SS}}$$

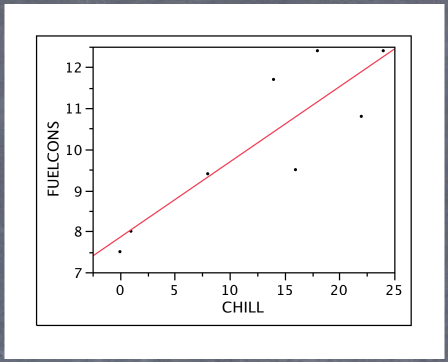
- Adjusted R-squared
  - Will not increase unless s^2 gets smaller
  - Difference from R^2 increases as k increases

$$\overline{R^2} = 1 - \frac{s^2}{\text{var}(y)}$$

Checking Assumptions

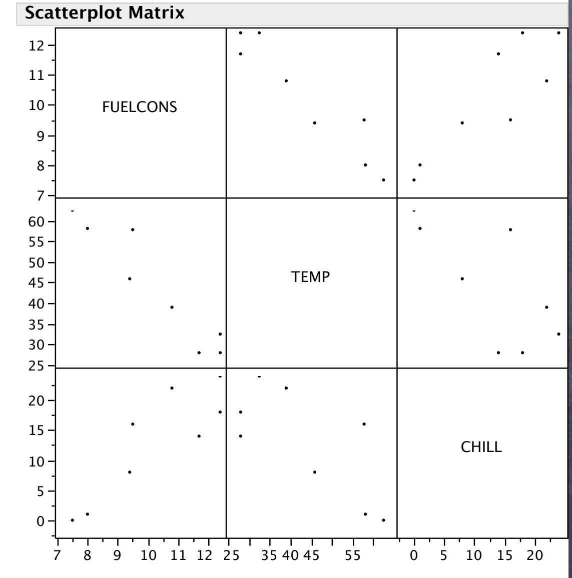
- Scatterplots of Y on X_1, Y on X_2
  - Data for fuel consumption (n = 8)





Data

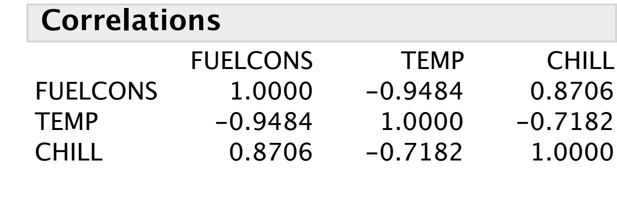
- Scatterplot matrix



y = weekly natural gas consumption

X_1 = average temperature

X_2 = chill index (wind, clouds, temp)

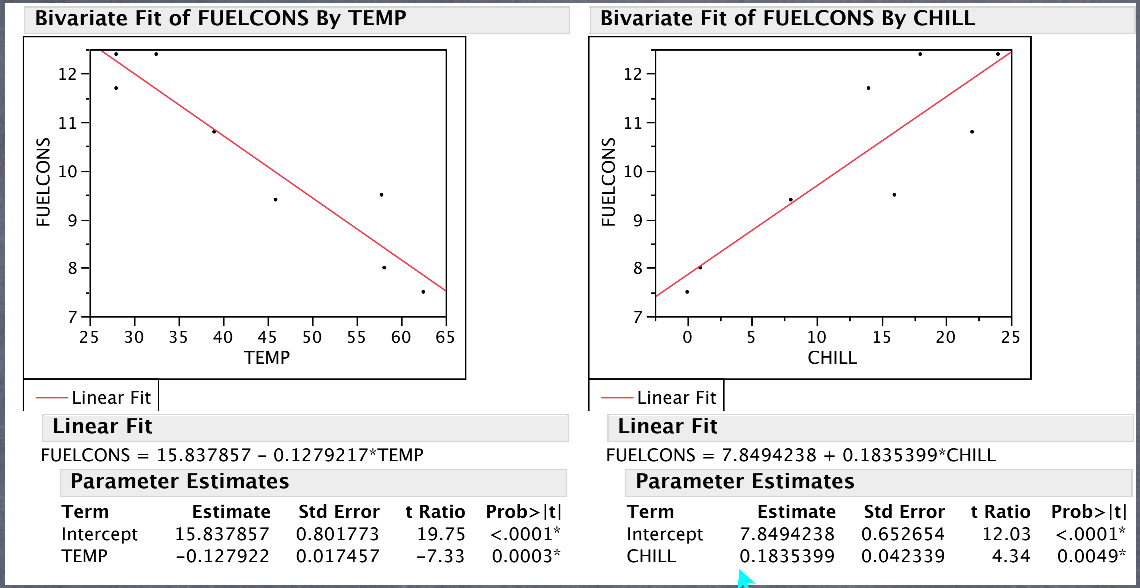


Correlations

	FUELCONS	TEMP	CHILL
FUELCONS	1.0000	−0.9484	0.8706
TEMP	−0.9484	1.0000	−0.7182
CHILL	0.8706	−0.7182	1.0000

Partial vs Marginal

- SRM



Bivariate Fit of FUELCONS By TEMP

Linear Fit

FUELCONS = 15.837857 - 0.1279217*TEMP

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	15.837857	0.801773	19.75	<.0001*
TEMP	-0.127922	0.017457	-7.33	0.0003*

Bivariate Fit of FUELCONS By CHILL

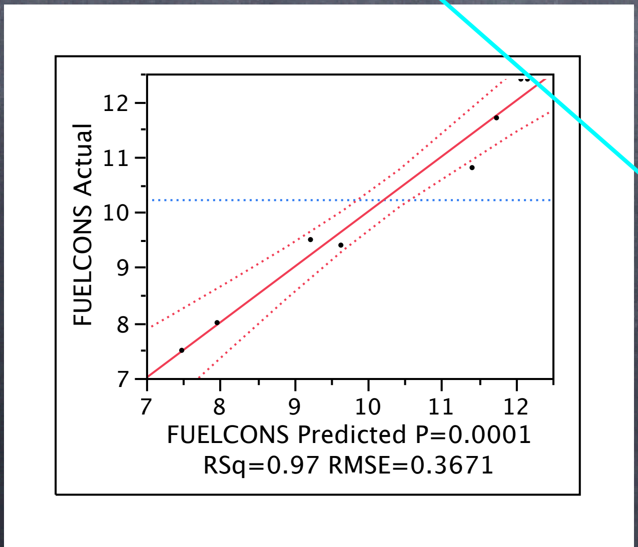
Linear Fit

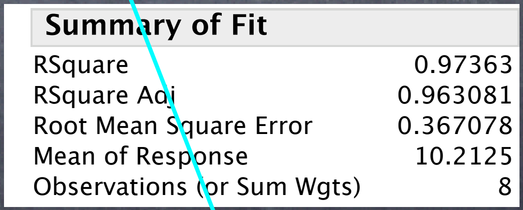
FUELCONS = 7.8494238 + 0.1835399*CHILL

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	7.8494238	0.652654	12.03	<.0001*
CHILL	0.1835399	0.042339	4.34	0.0049*

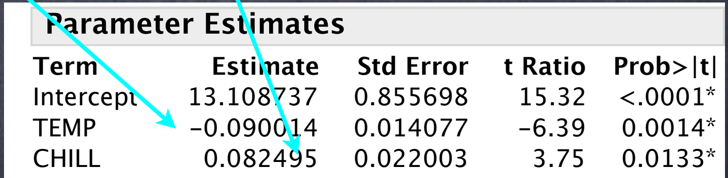
- MRM





Summary of Fit

RSquare	0.97363
RSquare Adj	0.963081
Root Mean Square Error	0.367078
Mean of Response	10.2125
Observations (or Sum Wgts)	8



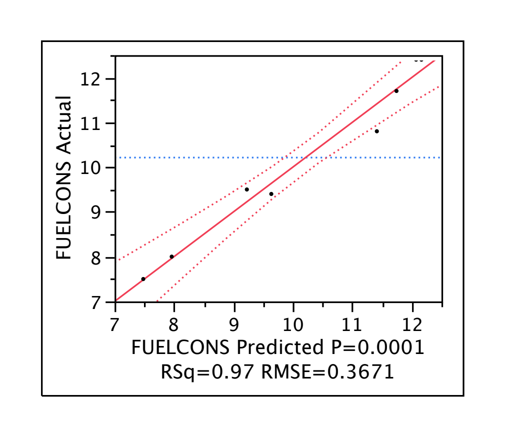
Parameter Estimates

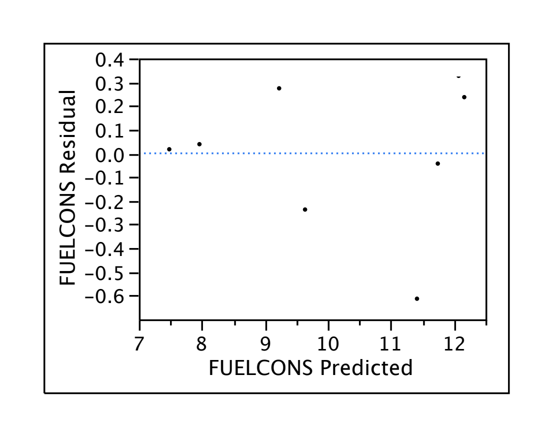
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	13.108737	0.855698	15.32	<.0001*
TEMP	-0.090014	0.014077	-6.39	0.0014*
CHILL	0.082495	0.022003	3.75	0.0133*

Slopes differ

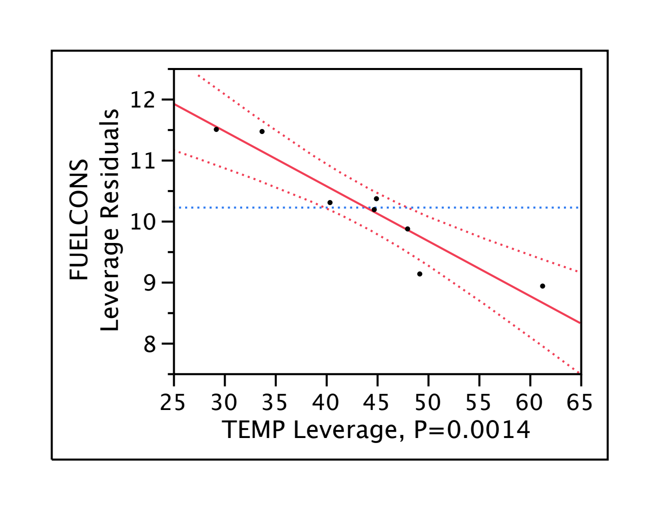
More Diagnostics

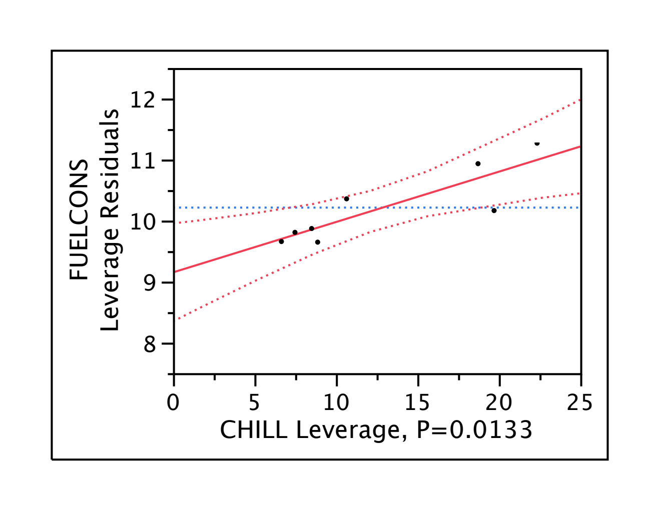
- Overall plots (MRM version of SRM scatterplots)





- Leverage plots (partial regression plots)
  - Simple regression view of MR slope, one for each slope





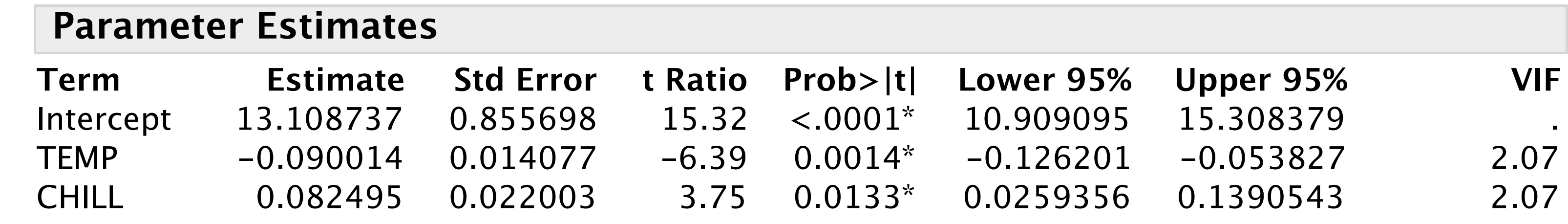
Inference

- Standard error of the slope is affected by correlation among explanatory variables
  - Variance inflation factor (Chap 5)

    Var(slope in MRM) ≈ Var(slope in SRM) VIF

$$\mathrm{Var}(b_j) = \frac{\sigma^2}{\sum_i (x_{ij} - \overline{x}_j)^2} \left(\frac{1}{1 - R^2_{X_j|X_{m\neq j}}} \right)$$

- Three equivalent methods for each estimated slope and the intercept
  - Confidence interval
  - t-statistic
  - p-value

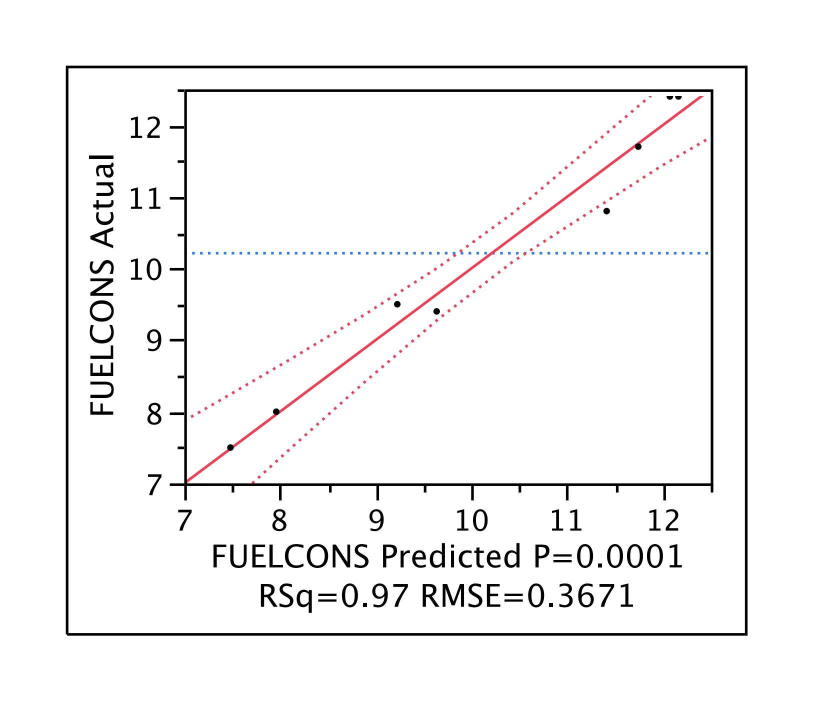


Parameter Estimates

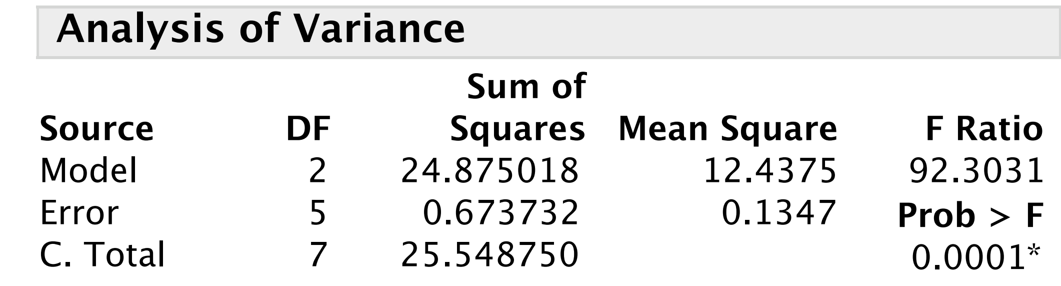
Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%	VIF
Intercept	13.108737	0.855698	15.32	<.0001*	10.909095	15.308379	.
TEMP	-0.090014	0.014077	-6.39	0.0014*	-0.126201	-0.053827	2.07
CHILL	0.082495	0.022003	3.75	0.0133*	0.0259356	0.1390543	2.07

Overall F Test

- Test both slopes simultaneously
  - H_0: $\beta_1 = \beta_2 = 0$
  - Ratio of variance explained to remaining variation
- Test of the size of R^2 statistic



$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$



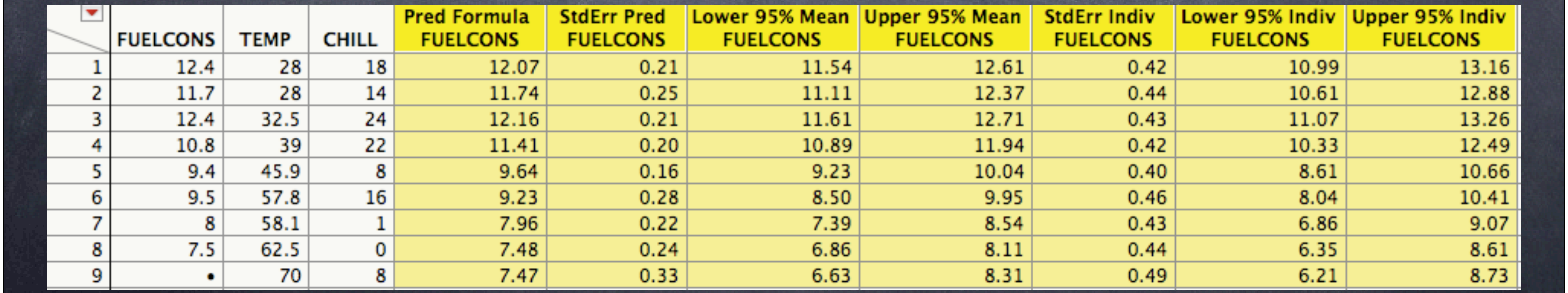
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	24.875018	12.4375	92.3031
Error	5	0.673732	0.1347	**Prob > F**
C. Total	7	25.548750		0.0001*

Prediction

- No simple plot
  - Extrapolation effect is more subtle
- Software is needed to identify extrapolation
  - Options in Fit Model to save various standard errors as well as prediction and confidence intervals
  - Add an extra row (before fitting) to get JMP to predict a new case

$\hat{y}$

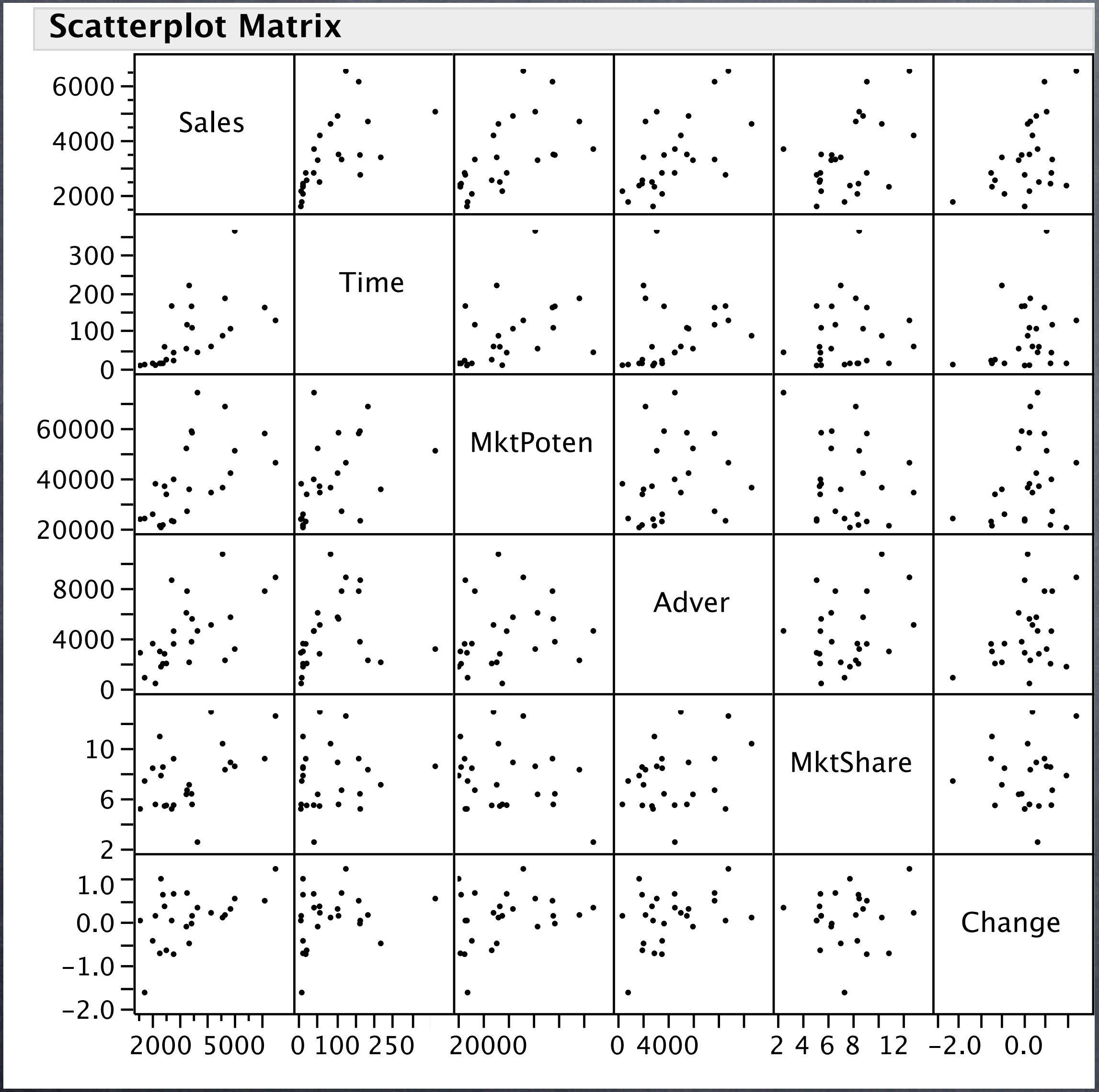


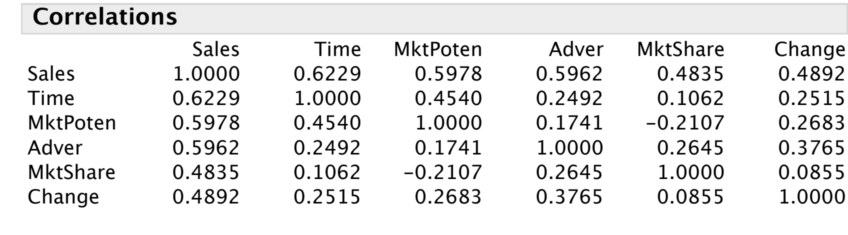
	FUELCONS	TEMP	CHILL	Pred Formula FUELCONS	StdErr Pred FUELCONS	Lower 95% Mean FUELCONS	Upper 95% Mean FUELCONS	StdErr Indiv FUELCONS	Lower 95% Indiv FUELCONS	Upper 95% Indiv FUELCONS
1	12.4	28	18	12.07	0.21	11.54	12.61	0.42	10.99	13.16
2	11.7	28	14	11.74	0.25	11.11	12.37	0.44	10.61	12.88
3	12.4	32.5	24	12.16	0.21	11.61	12.71	0.43	11.07	13.26
4	10.8	39	22	11.41	0.20	10.89	11.94	0.42	10.33	12.49
5	9.4	45.9	8	9.64	0.16	9.23	10.04	0.40	8.61	10.66
6	9.5	57.8	16	9.23	0.28	8.50	9.95	0.46	8.04	10.41
7	8	58.1	1	7.96	0.22	7.39	8.54	0.43	6.86	9.07
8	7.5	62.5	0	7.48	0.24	6.86	8.11	0.44	6.35	8.61
9	•	70	8	7.47	0.33	6.63	8.31	0.49	6.21	8.73

Sales Example

- Question
  - Evaluation of sales representatives
  - Response is annual company sales in territory
    - y measured in thousands of units
  - Data are a sample for n = 25 sales representatives
- Several explanatory variables
  - Time (months) with the company
  - Total sales of company and rivals in territory (potential)
  - Advertising expenditure in territory
  - Company's market share in prior four years
  - Change in company's market share

Initial Graphical Analysis

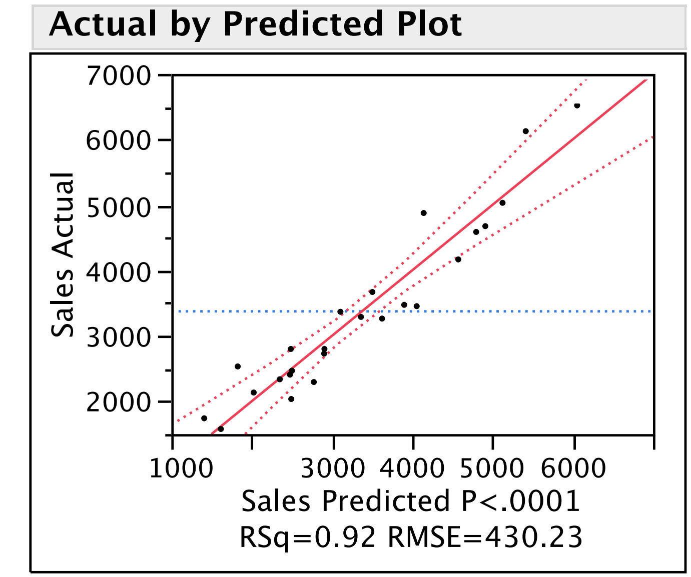

Scatterplot Matrix

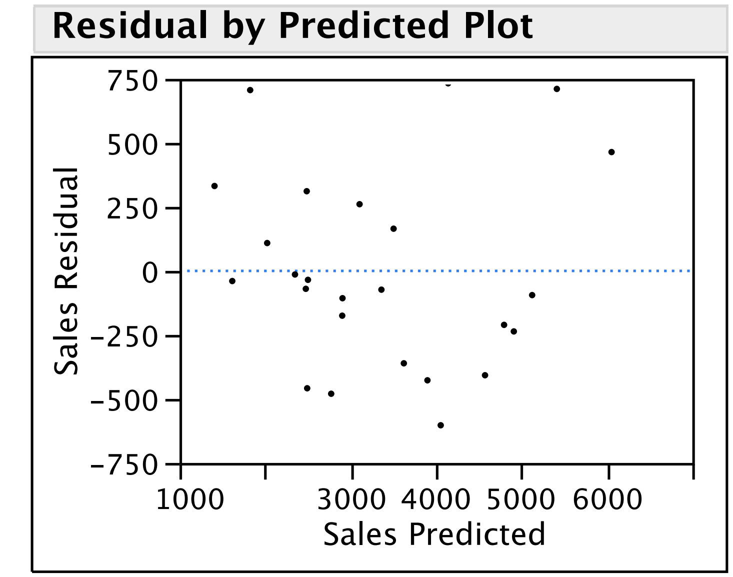

Correlations

	Sales	Time	MktPoten	Adver	MktShare	Change
Sales	1.0000	0.6229	0.5978	0.5962	0.4835	0.4892
Time	0.6229	1.0000	0.4540	0.2492	0.1062	0.2515
MktPoten	0.5978	0.4540	1.0000	0.1741	−0.2107	0.2683
Adver	0.5962	0.2492	0.1741	1.0000	0.2645	0.3765
MktShare	0.4835	0.1062	−0.2107	0.2645	1.0000	0.0855
Change	0.4892	0.2515	0.2683	0.3765	0.0855	1.0000

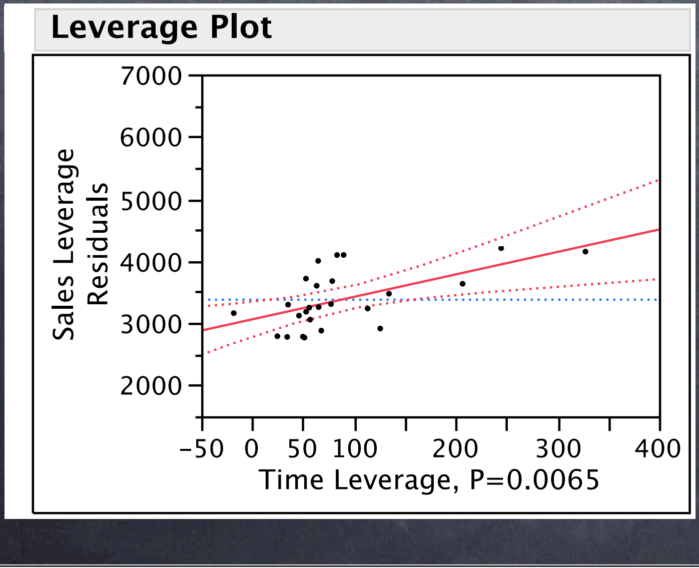
Multiple Regression

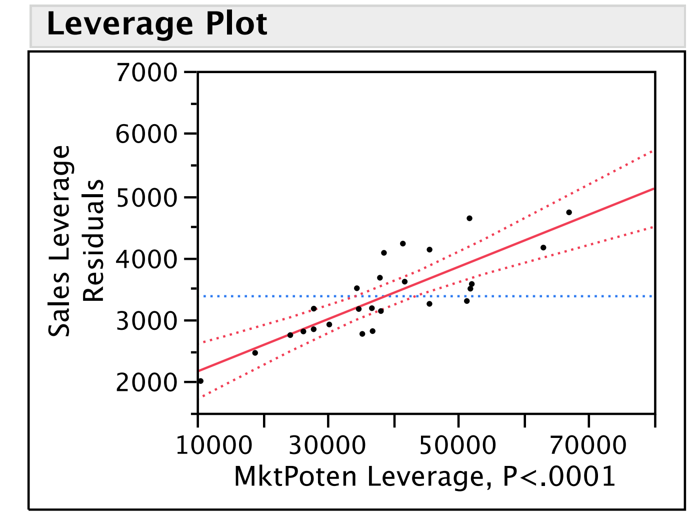
- Overall fit


Actual by Predicted Plot


Residual by Predicted Plot

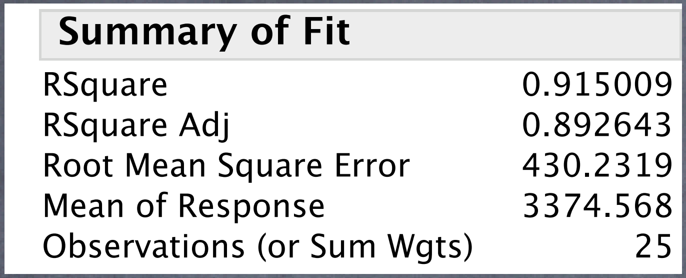
- Leverage plots


Leverage Plot


Leverage Plot

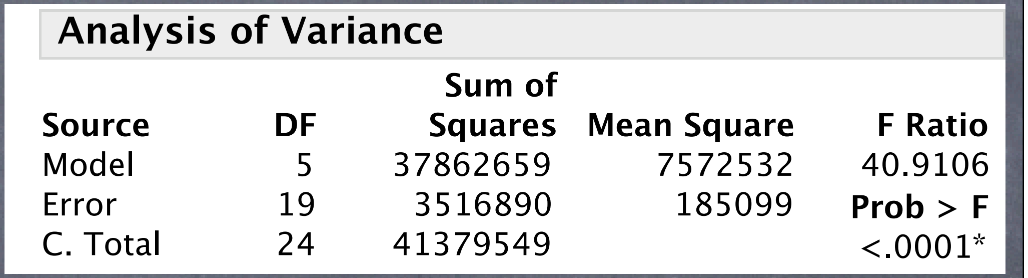
Model Summary

- Overall fit



Summary of Fit

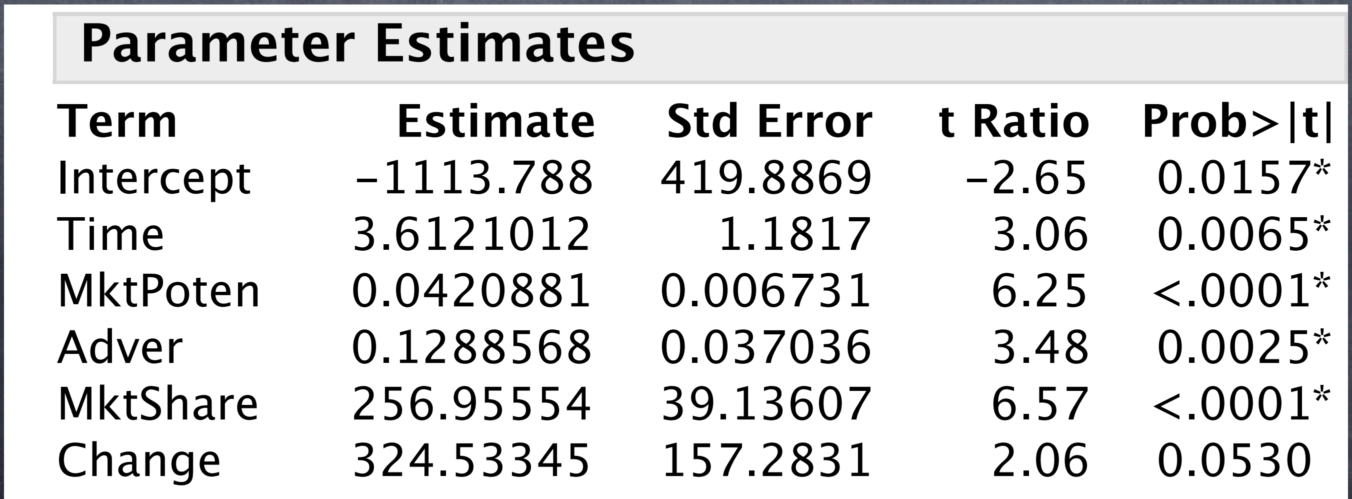
RSquare	0.915009
RSquare Adj	0.892643
Root Mean Square Error	430.2319
Mean of Response	3374.568
Observations (or Sum Wgts)	25



Analysis of Variance

		Sum of		
Source	**DF**	**Squares**	**Mean Square**	**F Ratio**
Model	5	37862659	7572532	40.9106
Error	19	3516890	185099	**Prob > F**
C. Total	24	41379549		<.0001*

- Individual estimates
  - Interpretation of these estimates?
  - Why linear? Implications of model are very strong.



Parameter Estimates

Term	**Estimate**	**Std Error**	**t Ratio**	**Prob>|t|**
Intercept	-1113.788	419.8869	-2.65	0.0157*
Time	3.6121012	1.1817	3.06	0.0065*
MktPoten	0.0420881	0.006731	6.25	<.0001*
Adver	0.1288568	0.037036	3.48	0.0025*
MktShare	256.95554	39.13607	6.57	<.0001*
Change	324.53345	157.2831	2.06	0.0530

Prediction

- Conditions for another rep (not one of these 25)
  - Sales were 3082

Time with company	85.42
Market potential	35,182.73
Advertising	7,281.65
Market share	9.64
Change in share	0.28

- Prediction results
  - Plug values for explanatory variables into equation
  - Prediction $\hat{y} = 4182$
  - Confidence interval for mean 3884.9 to 4478.6
  - Prediction interval for rep 3233.6 to 5129.9
  - Benchmarking implication: How is this rep doing?

Summary

- Multiple Regression Model (MRM)
- Estimators partial (MRM) vs marginal (SRM)
- Assumptions leverage plots
- Inference F-test
- Prediction Software